INFORMATION AND CONTROL 8, 80–92 (1965)

The Effect of Sequential Decision Feedback on Communication over the Gaussian Channel

ANDREW J. VITERBI

University of California, Los Angeles, California

INTRODUCTION

Considerable attention has been devoted over the past decade to communication systems employing a feedback channel to inform the transmitter concerning the state of the received information (Green, 1961). Two categories of feedback systems have been considered: information feedback wherein the receiver relays all or part of the received information back to the transmitter and the transmitter then attempts to correct the errors; and decision feedback wherein the receiver simply requests retransmission of questionable bits or words. The latter class of systems generally utilizes error detection procedures in deciding whether or not to request a retransmission. Most cases considered are for the binary symmetric forward channel and a noiseless or error-free feedback channel.

In this paper we consider coherent communication over a forward channel perturbed by additive white Gaussian noise and a noiseless feedback channel. We justify the latter partly as a means of determining the maximum effect of decision feedback for the Gaussian channel under ideal conditions and partly because for certain applications the feedback transmitter is several orders of magnitude more powerful than the forward transmitter, all other parameters of the two channels being the same. Then, particularly when the transmission rate in the reverse direction is much smaller than that in the forward direction, as will be the case for coded forward transmission, we may assume the feedback channel to be virtually error-free.

The feedback strategy which we employ may be termed *sequential decision feedback*. The receiver observes and operates on the received signal corresponding to a given bit or word until it is reasonably certain, according to some criterion, of being able to make the correct decision.

It then transmits a feedback signal which instructs the transmitter to begin sending the signal corresponding to the next bit or word. We also take into account the transmission path delay. The analysis of the system is based on sequential decision theory, which was originated by Wald (1947) and has been applied to a large class of problems including the related radar detection problem. We shall first consider binary uncoded transmission over the Gaussian channel and show that the ideal feedback channel affords a saving of approximately a factor of four in the energy-to-noise ratio relative to the conventional one-way system for the same error probability. We then turn to M'ary transmission wherein each transmitted signal corresponds to a sequence of $\log_2 M$ information bits and obtain bounds on the error probability when the transmitted signals are orthogonal. The bounds are of the nature of the exponential bounds of Fano (1961) for the one-way channel except that the negative exponent is considerably increased for all rates up to channel capacity.

BINARY COMMUNICATION

We consider first the case in which the information bits are transmitted by the signals $s_0(t)$ and $s_1(t)$ corresponding to "0" and "1", respectively, which are equally probable a priori. The decision to stop transmitting a given bit will be made when the probability of correctly identifying the bit has reached $1 - \epsilon$.

It is easily shown that the probability that signal $s_0(t)$ was transmitted, after the received waveform $y(t)$ has been observed for t seconds, is

$$P_0(t) = \frac{\exp (2Z_0(t)/N_0)}{\exp (2Z_0(t)/N_0) + \exp (2Z_1(t)/N_0)}$$

where

$$Z_i(t) = \int_0^t y(t)s_i(u)\, du \qquad (i = 0, 1)$$

and N_0 is the one-sided noise spectral density of the noise process $n(t)$ and the probability that $s_1(t)$ was transmitted, $P_1(t) = 1 - P_0(t)$. Thus when either $P_0(t)$ or $P_1(t)$ reach the value $1 - \epsilon$ for the first time the transmission is ended by a feedback signal. The problem then is to determine the expected time to reach a decision and to compare this with the fixed time required in a one-way system to ensure the same proba-

bility of error on the average. An exact solution will be obtained in the special case which is of greatest interest.

Let the signals have constant envelopes of the same magnitude but opposite signs, i.e.

$$s_0(t) = \sqrt{2S} \cos \omega t$$
$$s_1(t) = -\sqrt{2S} \cos \omega t$$

where ω is the carrier frequency. Then, assuming that $\omega t \gg 1$ so that we may neglect the double frequency term, when $s_0(t)$ is transmitted,

$$Z_0(t) = St + \sqrt{2S} \int_0^t n(u) \cos \omega u \, du = -Z_1(t) \qquad (1)$$

Thus for white Gaussian noise $Z_0(t)$ is a Gaussian variable of mean St and variance $N_0St/2$. In this case

$$P_0(t) = \frac{\exp (2Z_0(t)/N_0)}{\exp (2Z_0(t)/N_0) + \exp - (2Z_0(t)/N_0)}$$

When $s_1(t)$ is transmitted, the same is true with $Z_0(t)$ replaced by $Z_1(t) = -Z_0(t)$ and $P_0(t)$ replaced by $P_1(t)$. Thus it is sufficient to compare continuously the output of a correlator or matched filter, $Z_0(t)$, with the fixed constants $\pm a$, where a is a positive constant related to the required error probability, ϵ, as follows:

$$1 - \epsilon = \frac{\exp (2a/N_0)}{\exp (2a/N_0) + \exp - (2a/N_0)}. \qquad (2)$$

and as soon as $| Z_0(t) | = a$ the transmission is ended. The process $Z_0(t)$ is a Markov process (cf. Eq. (1)). Darling and Siegert (1953) have shown that the expected time, T, for a Markov process to reach the level $+a$ or $-a$ starting from an arbitrary value $Z(-a < Z < a)$ may be obtained by solving the differential equation

$$\frac{Bd^2T}{2dZ^2} + A \frac{dT}{dZ} = -1$$

with boundary conditions $T(a) = T(-a) = 0$ where

$$A = \lim_{\Delta t \to 0} \left\{ \frac{1}{\Delta t} E[\Delta Z_0(t)] \right\} = S$$
$$B = \lim_{\Delta t \to 0} \left\{ \frac{1}{\Delta t} E[\Delta Z_0^2(t)] \right\} = \frac{N_0 S}{2}$$

Thus the expected time to reach $\pm a$ from an initial value of zero is the solution for $Z = 0$. This is readily obtained to be

$$\bar{T} = T(0) = \frac{a}{S} \tanh \left(\frac{2a}{N_0} \right) < \frac{a}{S} \tag{3}$$

Equations (2) and (3) relate the expected time to the error probability. If we use the upper bound in (3), which is a close approximation for large a/N_0, we have

$$P_E = \epsilon < \frac{\exp - (4S\bar{T}/N_0)}{1 + \exp - (4S\bar{T}/N_0)} \tag{4}$$

This should be compared with the well-known result for the average error probability for a fixed time one-way channel:

$$\bar{P}_E = \operatorname{erfc} \sqrt{\frac{2S\bar{T}}{N_0}} < \frac{\exp - (S\bar{T}/N_0)}{\sqrt{4\pi S\bar{T}/N_0}} . \tag{5}$$

where the upper bound is a close approximation for large energy-to-noise ratios. Thus we note that for small error probabilities the feedback channel affords a conservation of almost a factor of 4 in the average energy-to-noise ratio. The exact expression of (5) is compared to the bound of (4) in Fig. 1.

A characteristic of this sequential decision strategy is that the error probability is constant for all transmissions, while the transmission time is a random variable. For one-way channels, on the other hand, the transmission time is fixed but the error probability varies from one transmission to the next, and we generally are satisfied with a knowledge of its ensemble average.

M'ARY COMMUNICATION

We consider now the case of coded transmission wherein words consisting of sequences of $\log_2 M$ bits are transmitted by means of one of M possible signals $s_{(i)}(t) (i = 0, 1, 2, \cdots M - 1)$. The same decision strategy for binary communication generalizes to the formation of the M statistics

$$P_i(t) = \frac{\exp (2Z_i(t)/N_0)}{\sum_{j=0}^{M-1} \exp (2Z_j(t)/N_0)} \qquad i = 0, 1, \cdots M - 1 \tag{6}$$

where

$$Z_i(t) = \int_0^t y(u) s_{(i)}(u)\, du$$

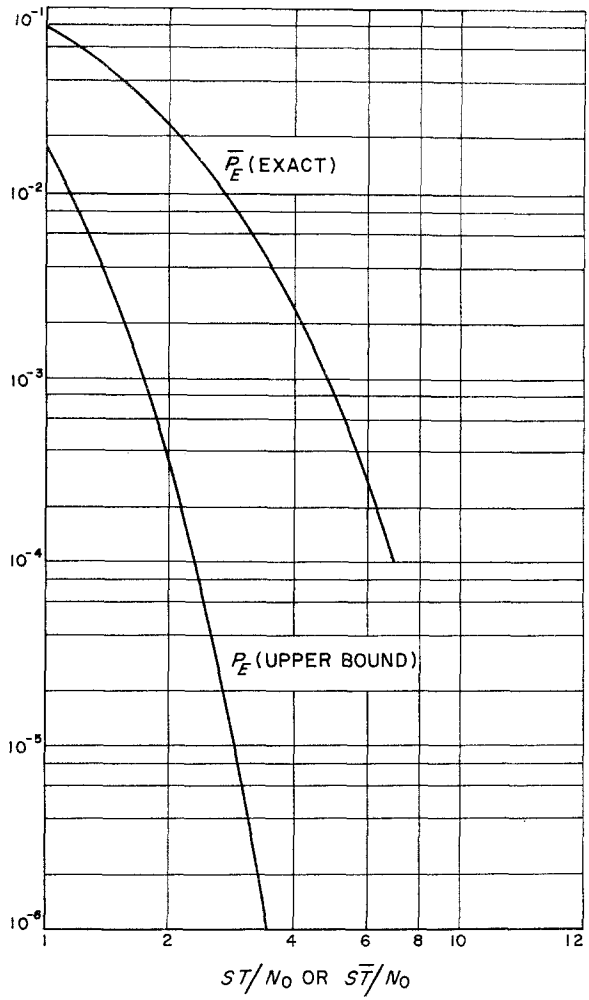

FIG. 1. Comparison of error probabilities for binary communications with one-way and feedback channels.

As soon as $P_i(t)$ reaches the value $1 - \epsilon$ for some i the transmission is terminated. Unfortunately, since the random processes $P_i(t)$ are not Gaussian we are unable to compute the expected time for this event.

Instead we will consider the suboptimal strategy of comparing the M statistics $Z_i(t)$ with a threshold level, a, and terminating the transmission

as soon as $Z_j(t) = a$ for some j and deciding in favor of the corresponding signal. With this strategy we must not only compute the expected time to termination but also the probability that a $Z_j(t)$ other than that corresponding to the correct signal crosses the level first. In this case both the time to termination and the error probability are random variables and we shall compute the expectations of both.

In order to proceed we must restrict our attention to the case in which the $Z_i(t)$ are independent random processes. Strictly speaking this is physically impossible since it requires that the signals $s_{(k)}(t)$ be mutually orthogonal for all t. However, we can approach this condition as closely as desired provided the system is not bandwidth limited. For example, if we choose $s_{(k)}(t) = \sqrt{2S}\cos(\omega + k\omega')t (k = 0, 1, 2, \cdots M - 1)$ we find that the pertinent moments of $Z_k(t)$ under the hypothesis that $s_{(i)}(t)$ was transmitted are

$$\begin{aligned} E[Z_k(t)] &= St\frac{\sin(i-k)\omega' t}{(i-k)\omega' t} \\ \mathrm{Var}[Z_k(t)] &= \frac{N_0 St}{2} \\ \mathrm{Cov}[Z_k(t)Z_m(t)] &= \frac{N_0 St}{2}\left[\frac{\sin(k-m)\omega' t}{(k-m)\omega' t}\right] \end{aligned} \tag{7}$$

Thus since for $k \neq m$, the magnitude of the covariance is bounded by

$$|\operatorname{Cov}[Z_k(t)Z_m(t)]| < \frac{N_0 S}{2|k-m|\omega'}$$

while the expectation of $Z_k(t)$ for $k \neq i$

$$|E[Z_k(t)]| < \frac{S}{|i-k|\omega'}.$$

if we choose ω' sufficiently large that $N_0 S/\omega' \ll 1$ and $S/\omega' \ll 1$ then the processes are essentially independent and the expectation of all processes except $Z_i(t)$ are essentially zero.

Without loss of generality let us assume that $s_{(0)}(t)$ was transmitted. Then $Z_0(t)$ is the Markov process described by Eq. (1). By solving a Fokker-Planck equation for this process with appropriate boundary condition the probability density function for the first level crossing time can be shown to be (Darling and Siegert, 1953; Feller, 1950)

$$p_0(t) = \frac{a}{(\pi N_0 S)^{1/2} t^{3/2}} \exp - \frac{(a - St)^2}{N_0 St} \tag{8}$$

while for the other processes, assuming $E[Z_k(t)] = 0$, we have

$$p(t) = p_k(t) = \frac{a}{(\pi N_0 S)^{1/2} t^{3/2}} \exp - \left(\frac{a^2}{N_0 St}\right) \quad k = 1, 2, \cdots M - 1 \tag{9}$$

The expected time for $Z_0(t)$ to cross[1] the level a is, therefore,

$$T_0 = \int_0^\infty t p_0(t)\, dt = \frac{a}{S} \tag{10}$$

It is possible that one of the other processes $Z_k(t)$ crosses the level a before T_0. Thus the expected time to termination is

$$\bar{T} < T_0 = a/S \tag{11}$$

However, when the error probability is small, this will naturally be a low probability event and the bound (11) is closely approximated.

The probability of error is just the probability that $Z_k(t)$ reaches a before $Z_0(t)$ for some $k \neq 0$ and since the processes are taken to be independent this is

$$\begin{aligned} \bar{P}_E &= 1 - \int_0^\infty p_0(t) \left[\int_t^\infty p(u)\, du\right]^{M-1} dt \\ &< M \int_0^\infty p_0(t)\, dt \int_0^t p(u)\, du \end{aligned} \tag{12}$$

since $(1 - x)^{M-1} > 1 - Mx$ for $x < 1$. Then using the bounds erfc $x <$ exp $- (x^2/2)$ and (11) we obtain

$$\bar{P}_E < \sqrt{2}\, M \exp \frac{2S\bar{T}}{N_0} (1 - \sqrt{2}). \tag{13}$$

Since in $\bar{T}$ sec on the average $\log_2 M$ bits or ln M nats are transmitted, the average rate is $\bar{R} =$ ln M/T nats/sec, while for a white Gaussian channel with no bandwidth limitations the channel capacity is $C =$ S/N_0 nats/sec. Therefore:

$$\frac{C}{\bar{R}} = \frac{S\bar{T} \ln M}{N_0} \tag{14}$$

[1] Note that this is a one-sided barrier crossing problem as distinguished from the two-sided barrier problems considered previously for binary antipodal signals.

and (13) becomes

$$\bar{P}_E < \sqrt{2}\, M^{-(C/\bar{R})[2(\sqrt{2}-1)-(\bar{R}/C)]} \tag{15}$$

This bound is adequate for small $\bar{R}/C$ but the exponent goes to zero at $\bar{R}/C = 2(\sqrt{2} - 1 \approx 0.828$, while we know that even for the one-way channel, it can be made negative for all $\bar{R}/C < 1$.

An improved bound for large $\bar{R}/C$ is obtained by the following argument. Suppose that if no decision were reached by a given time τ, the transmission would be arbitrarily ended without a decision. Then the error probability would be

$$\bar{P}_E = P_1 + P_2 \tag{16}$$

where

$$\begin{aligned} P_1 &= \text{Prob (first level crossing of } Z_0(t) \text{ occurs after } \tau \text{ sec)} \\ &= \int_\tau^\infty p_0(t)\, dt \end{aligned} \tag{17}$$

and

$$\begin{aligned} P_2 &= \text{Prob (first level crossing of } Z_0(t) \text{ occurs within } \tau \text{ sec and} \\ &\quad Z_j(t) \text{ crosses level before } Z_0(t) \text{ for some } j \neq 0) \\ &< M \int_0^\tau p_0(t)\, dt \int_0^t p(u)\, du \\ &< 2M\bar{T} \left(\frac{S}{\pi N_0}\right)^{1/2} \int_0^\tau \exp - \frac{(S/N_0)[2(\bar{T}^2/t) - 2\bar{T} + t]}{t^{3/2}}\, dt \end{aligned} \tag{18}$$

Equation (18) follows from the same argument which led to (13) and if we let $\tau = \infty$ (18) reduces to (13) and (17) vanishes. P_1 and P_2 can be further bounded by Chernov distribution tilting methods (Fano, 1961, p. 257). These are obtained in the Appendix as

$$P_1 < \exp - \frac{S}{N_0}\left(\tau + \frac{\bar{T}^2}{\tau} - 2\bar{T}\right) \qquad (\bar{T} < \tau) \tag{19}$$

$$P_2 < \sqrt{2}\, M \exp - \frac{S}{N_0}\left(\tau + \frac{2\bar{T}^2}{\tau} - 2\bar{T}\right) \qquad (\bar{T} > \tau/2) \tag{20}$$

To find the value of τ which minimizes this requires solution of a transcendental equation. However, a reasonable approximation to the

minimum is obtained by setting $P_2 = \sqrt{2}\, P_1$. This yields

$$\frac{\tau}{\bar{T}} = \frac{S\bar{T}}{N_0 \ln M} = \frac{C}{\bar{R}} \tag{21}$$

and an upper bound

$$\bar{P}_E < (1 + \sqrt{2}) M^{-(C/\bar{R})[(\bar{R}/C)+(C/\bar{R})-2]} \tag{22}$$

The exponent is negative for all $\bar{R}/C < 1$.

The conditions on the Chernov bounds

$$\tau/\sqrt{2} < \bar{T} < \tau$$

become, upon replacing a by $S\bar{T}$ and using (21),

$$1/\sqrt{2} < \bar{R}/C < 1$$

Combining the results (15) and (22) we have

$$P_E < K M^{-(C/\bar{R})\alpha} \text{ (with feedback)} \tag{23}$$

where

$$K = \begin{cases} \sqrt{2} & (0 < \bar{R}/C \leqq 1\sqrt{2}) \\ 1 + \sqrt{2} & (1/\sqrt{2} \leqq \bar{R}/C < 1) \end{cases}$$

$$\alpha = \begin{cases} 2(\sqrt{2} - 1) - \bar{R}/C & (0 \leqq \bar{R}/C \leqq 1/\sqrt{2}) \\ (\sqrt{C/\bar{R}} - \sqrt{\bar{R}/C})^2 & (1/\sqrt{2} \leqq \bar{R}/C < 1) \end{cases}$$

α is shown as a function of $\bar{R}/C$ in Fig. 2. Note that the two functions, α are tangent at the point $1/\sqrt{2}$. $\bar{R}/C$ is related to the average energy-to-noise ratio by (14). These results are to be compared with similar bounds for the one-way channel first obtained by Fano (1961)

$$P_E < K' M^{-(C/R)\alpha'} \qquad \text{(one-way channel)}$$

where

$$\alpha' = \begin{cases} \tfrac{1}{2} - (R/C) & 0 < R/C \leqq \tfrac{1}{4} \\ (1 - \sqrt{R/C})^2 & \tfrac{1}{4} \leqq R/C < 1 \end{cases} \tag{24}$$

α' is shown as a function of R/C in Fig. 2. K' is a rather complicated function of R/C. However, if Chernov bounds are used instead of the Stirling bounds used by Fano, one obtains $K' < 2$. Furthermore, Zetter-

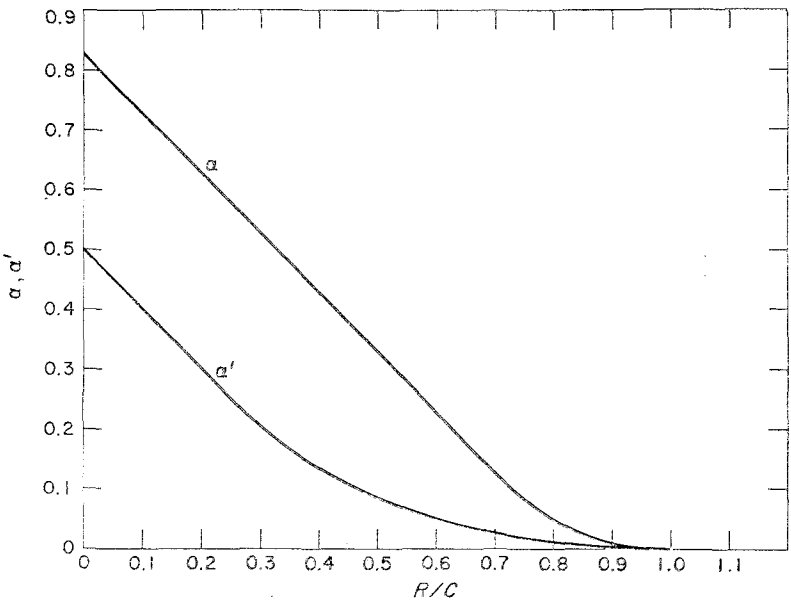

FIG. 2. Exponents of error probability bounds for one-way and feedback channels.

berg (1961) has shown that a lower bound on the error probability is

$$\bar{P}_E > K'' M^{-(C/R)\alpha'}$$

where α' is as given in (24) and K'' is a rather complicated function of R/C and $\log_M K'' \to 0$ as $M \to \infty$.

Thus since the exponent dominates the expression for large M, and α' is the exponent for both upper and lower bounds, comparison of α with α' yields considerable information on the improvement afforded by the feedback channel. In Fig. 3 is shown the ratio α/α', which reaches a maximum of 4.8 at $R/C \approx 0.7$ and remains above 4 up to capacity. For low rates the ratio falls below 2. This is somewhat surprising particularly in view of the fact that we showed in the previous section that with an optimal decision strategy, binary communication (which being uncoded has naturally a low rate for low error probabilities) showed an improvement of nearly four for large energy-to-noise ratios. Most likely this is a consequence of the suboptimal strategy used here.

It is also interesting to note that when we impose the time limitation

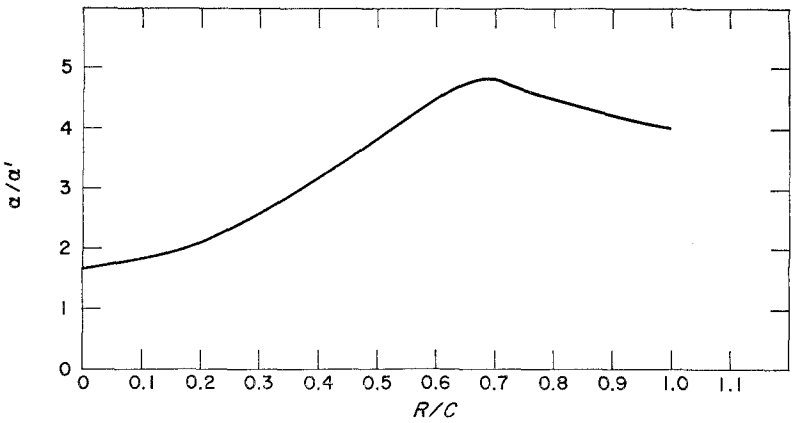

FIG. 3. Ratio of exponents of error probability bounds for one-way and feedback channels.

to obtain the bound (22), the ratio of the maximum time allowed to the average time without the restriction is just the ratio of capacity to rate (cf. (21)). Furthermore, the temporal threshold τ used in the argument which led to (22) is reminiscent of the amplitude threshold and the corresponding argument which may be used to derive the expression for the exponent for one-way channels, in the range $\frac{1}{4} < R/C < 1$ given in (24).

CONCLUSIONS

The results of (4), (5), (23), and (24) shown graphically in Figs. 1, 2, and 3 indicate that noiseless sequential decision feedback can produce a significant improvement in performance. Specifically for binary uncoded communication the error probability is raised to nearly the fourth power for high energy-to-noise ratio, or equivalently for a fixed low error probability the energy-to-noise ratio may be reduced by a factor of four. For coded M'ary communications using a somewhat inferior strategy, the corresponding factor varies from approximately two to five depending on the transmission rate.

From a practical standpoint the method is limited by two principal drawbacks: any error in the feedback channel consisting of either a false alarm or a failure to detect the feedback command to terminate the present transmission will cause a synchronization error in all the future received data. It should be noted, however, that the feedback

channel is used to transmit only synchronization information, while the forward channel is basically asynchronous. Thus, as is always the case in a communication system, the synchronization channel must be superior to the data channel.

The other limitation involves the propagation delay as well as the delay required for the feedback command. We cannot require the transmitter to determine the exact instant of time at which the command was issued for this would require an infinite rate in the feedback channel. On the other hand, if we restrict the feedback command pulse to be sent during certain fixed intervals, the transmitter must simply determine whether a command was issued in a particular interval. Then if δ is the two-way propagation delay and γ is the interval between possible command pulses, the average time per transmitted word, $\bar{T}$, is increased by at most $\delta + \gamma$. Consequently $\bar{R}$ must be replaced by $\bar{R}\,[\bar{T}/(\bar{T} + \delta + \gamma)]$ in all the bounds. Thus, for example, when $\delta + \gamma = \bar{T}$, the performance indicated above is degraded by a factor of two.

APPENDIX. CHERNOV BOUNDS (19) AND (20)

The probability distribution[2] $P(\tau) \leqq \int_0^\tau p(t)\,dt$ satisfies the inequality

$$P(\tau) \leqq \exp - [s_0\gamma'(s_0) - \gamma(s_0)] \qquad (s_0 \leqq 0) \qquad \text{(A.1)}$$

where

$$\gamma(s) = \ln g(s),$$

$g(s)$ is the moment generating function of $p(\tau)$, and s_0 is the solution of the equation $\gamma'(s_0) = \tau$, provided the derivative exists and is continuous. Similarly, for the complement of a probability distribution,

$$1 - P(\tau) \leqq \exp - [s_0\gamma'(s_0) - \gamma(s_0)] \quad (s_0 \geqq 0) \qquad \text{(A.2)}$$

P_1 of (17) is the complement of a distribution. The moment generating function is

$$g_1(s) = \int_0^\infty p_0(t)e^{st}\,dt \exp \frac{2S\bar{T}}{N_0}\left[1 - \left(1 - \frac{sN_0}{S}\right)^{1/2}\right]$$

Thus

$$\gamma_1'(s) = \bar{T}\left(1 - \frac{sN_0}{S}\right)^{-1/2}$$

[2] Cf. Fano (1961), p. 257.

and

$$s_0 = \frac{S}{N_0}\left[1 - \left(\frac{\bar{T}}{\tau}\right)^2\right] \geqq 0$$

Insertion of these in (A.2) yields (19).

The bound on P_2 of (18) is not as such a distribution function but we can transform it into one by normalizing by its value for $\tau = \infty$ which is given by (12). Thus

$$g_2(s) = \exp \frac{2\sqrt{2}S\bar{T}}{N_0}\left[1 - \left(1 - \frac{sN_0}{S}\right)^{1/2}\right]$$

and

$$\gamma_2'(s) = \sqrt{2}\,\bar{T}\left(1 - \frac{sN_0}{S}\right)^{-1/2}$$

while

$$s_0 = \frac{S}{N_0}\left[1 - 2\left(\frac{\bar{T}}{\tau}\right)^2\right] \leqq 0.$$

Inserting these in (A.1) and multiplying by (13) we obtain (20).

RECEIVED: April 6, 1964.

REFERENCES

DARLING, D. A., AND SIEGERT, A. J. F. (1953), The first-passage time problem for a continuous Markov process. *Ann. Math. Statist.* **24**, 624–639.

FANO, R. M. (1961), "Transmission of Information." MIT Press and Wiley, New York.

FELLER, W. (1950), "An Introduction to Probability Theory and Its Applications." Wiley, New York.

GREEN, P. E., JR., (1961), Feedback communication systems. In "Lectures on Communication System Theory, E. J. Baghdady, ed. McGraw-Hill, New York.

WALD, A. (1947), "Sequential Analysis." Wiley, New York.

ZETTERBERG, L.-H. (1961), Data Transmission over a noisy Gaussian channel. *Kungl. Tekn. Högskol. Handl. Stockholm*, No. 184, 1–87.